Text2Story Lusa: A Dataset for Narrative Analysis in European Portuguese News Articles

Sérgio Nunes[1], Alípio Jorge[1], Evelin Amorim[1], Hugo Sousa[1], António Leal[2], Purificação Silvano[2], Inês Cantante[2], Ricardo Campos[3]

[1]INESC TEC and University of Porto
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal

[2]CLUP and University of Porto
Via Panorâmica, 4150-564 Porto, Portugal

[3]INESC TEC and University of Beira Interior and Ci2 - Smart Cities Research Centre (IPTomar)
Covilhã, Portugal

sergio.nunes@fe.up.pt, amjorge@fcc.up.pt, evelin.f.amorim@inesctec.pt, hugo.o.sousa@inesctec.pt, jleal@letras.up.pt, msilvano@letras.up.pt, cantante.ines@gmail.com, ricardo.campos@ubi.pt

Abstract

Narratives have been the subject of extensive research across various scientific fields such as linguistics and computer science. However, the scarcity of freely available datasets, essential for studying this genre, remains a significant obstacle. Furthermore, datasets annotated with narratives components and their morphosyntactic and semantic information are even scarcer. To address this gap, we developed the Text2Story Lusa datasets, which consist of a collection of news articles in European Portuguese. The first datasets consists of 357 news articles and the second dataset comprises a subset of 117 manually densely annotated articles, totaling over 50 thousand individual annotations. By focusing on texts with substantial narrative elements, we aim to provide a valuable resource for studying narrative structures in European Portuguese news articles. On the one hand, the first dataset provides researchers with data to study narratives from various perspectives. On the other hand, the annotated dataset facilitates research in information extraction and related tasks, particularly in the context of narrative extraction pipelines. Both datasets are made available adhering to FAIR principles, thereby enhancing their utility within the research community.

Keywords: text dataset, manual annotations, natural language processing, narrative extraction, news articles, European Portuguese

1. Introduction

The growing interest in narrative understanding, either from a linguistic or computational perspective, has led to a search for data, which can be a complex endeavor, as well as for annotated corpora. These corpora are used for applying natural language processing techniques to extract, summarize, and recreate narratives (Santana et al., 2023). Data acquisition can present significant challenges for two primary reasons: the absence of readily available datasets (Ide, 2017) and copyright concerns (Zeldes, 2017). Access to annotated corpora is also limited, with few available options, many of which are either paid or incomplete in terms of the necessary information for comprehensive research and narrative element extraction.

In the context of the Text2Story[1] project, we were able to overcome these obstacles and created the Text2Story Lusa dataset and the Text2Story Lusa Annotated dataset, aimed at facilitating research in linguistics and computer science. This resource is published as two independent datasets: the Text2Story Lusa (Nunes et al., 2023) and the Text2Story Lusa Annotated Corpus (Silvano et al., 2023).

The Text2Story Lusa dataset consists of 357 full-text news articles written in European Portuguese, sourced from Lusa[2], the largest Portuguese news agency. The Text2Story Lusa Annotated Corpus comprises manual dense annotations for a subset of 117 articles from the full collection. These annotations offer additional linguistic information, which includes morphosyntactic and semantic details about the main elements that compose a narrative. In total, there are over 7 thousand entity annotations (e.g., participants, events, times), more than 12 thousand link annotations (e.g., semantic roles, temporal, objectal), and over 31 thousand attributes defined for both entity and link annotations.

Both datasets were created with the primary motivation of advancing research and development in

[1] https://text2story.inesctec.pt

[2] https://www.lusa.pt

narrative annotation and narrative extraction within the fields of linguistics and computer science. The availability of language resources plays a crucial role in conducting research in these areas (Jorge et al., 2019). With the release of these datasets, we contribute to the collection of public datasets available for European Portuguese, a low-resourced language.

These datasets stand out among existing resources because they gather complete texts, in this case, news of a narrative nature, in contemporary European Portuguese, which can be used for different purposes in several areas. Furthermore, the annotated corpus is the first one that combines several parts of the ISO 24617-Language resource management-Semantic annotation framework (Ide et al., 2003), ensuring interoperability in a time when the Semantic Web and Linguistic Linked Data are growing and interoperability is critical for interpreting linguistic resources. In addition, the exhaustive and intricate annotation allows for characterizing the central elements of the narrative (i.e., participants, events, time, place, etc). Therefore, beyond narrative extraction, the Text2Story Lusa datasets have the potential to support various other tasks and challenges in the field of Natural Language Processing (NLP). For instance, researchers can leverage this resource to explore information extraction, relation extraction, co-reference resolution, and other related tasks and challenges in NLP.

A distinct feature of these resources is their manual curation, which has resulted in a high-quality selection of news texts rich in narrative structures, plus a dense semantic annotation for these structures (over 50 thousand annotations). Additionally, they are made available under the Creative Commons Attribution-NonCommercial 4.0 International license (CC BY-NC)[3], which facilitates use by the research community.

The remainder of this paper is organized as follows. Section 2 presents related contributions. Section 3 provides a detailed description of the design, collection process, manual annotation, and a general characterization of the Text2Story Lusa datasets. Section 4 lists several examples of applications leveraging these resources. Section 5 discusses ethical considerations related to data collection and use. Section 6 presents the conclusions and future work perspectives.

2. Related Work

Despite the proliferation of corpora in recent years across various languages, many consist of texts from different genres. Even when they include news articles, these may not necessarily exhibit a narrative nature, which is essential for research focused on narrative understanding. Furthermore, while many of these datasets include annotations, they often prove of little use for the study of narratives if the texts themselves are not narrative in nature. For instance, the Groningen Meaning Bank (GMB) (Basile et al., 2012; Bos et al., 2017) encompasses data from various genres, including news and narrative texts such as fables, alongside other genres. Despite its broad annotation, which includes morphological, syntactic, and semantic aspects, the inclusion of non-narrative texts limits its utility for narrative-focused research.

The Manually Annotated Sub-Corpus (MASC) (Ide et al., 2008) and Georgetown University Multilayer Corpus (GUM) (Zeldes and Simonson, 2016) are two other examples of annotated corpora with an assortment of several types of texts. On the other hand, OntoNotes (Hovy et al., 2006), for instance, only gathers English and Chinese newswires. Still, the data is not exclusively narrative-focused, nor does the annotation aim to capture narrative elements. The same applies to the Causal News Corpus (CNC) (Tan et al., 2022), which is only annotated with causal relations.

Regarding Portuguese, in the Brazilian variety, there is, for example, CST news with cross-document annotated relations (Cardoso et al., 2011) and a small corpus of 50 annotated news texts (Collovini et al., 2007) with discourse relations. Similarly, for European Portuguese, there are also a few corpora and resources focused solely on news. It is the case of the resource CETEMPúblico (Rocha and Santos, 2000). However, this resource only provides partial extracts from articles, making it impossible (by design) to reconstruct the complete text.

Another example is the FEUP News Corpus (n.a., 2016) and its associated annotations (Mendes and Hendrickx, 2021). While details about the text included in the FEUP News Corpus are limited, it represents a crawling of news articles from Portuguese online news media sources, scraped from the web. This collection comprises a heterogeneous set of web articles gathered without known criteria. The CINTIL-Corpus Internacional do Português (Barreto et al., 2006)[4] and the Reference Corpus of Contemporary Portuguese (Généreux et al., 2012)[5] also are made of news articles in European Portuguese. However, the full texts are unavailable, and the most recent data is from 2006. Moreover, and for the most part, the annotations concern morphosyntactic aspects.

The two Text2Story Lusa datasets we have cre-

[3] https://creativecommons.org/licenses/by-nc/4.0/

[4] https://portulanclarin.net/workbench/cintil-concordancer/

[5] http://gamma.clul.ul.pt/CQPweb/

ated feature a set of distinguishing characteristics not found in existing datasets. Firstly, they are shared under a clear legal protocol with the intellectual property holders. Secondly, they contain a uniform collection of news articles written in contemporary European Portuguese, a low-resourced language, and selected based on well-defined criteria to ensure the creation of a coherent narrative corpus. Thirdly, the Text2Story Lusa Annotated Corpus features a multilayer manual annotation (Silvano et al., 2021b; Leal et al., 2022), which captures the main narrative components. For these reasons, these datasets represent a unique and valuable resource.

3. Text2Story Lusa Datasets

In this section, we present the Text2Story Lusa datasets, first by describing the creation process in Subsection 3.1, then presenting the manual annotation in Subsection 3.2, and finally providing an overview of both resources with a characterization in Subsection 3.3.

3.1. Dataset Creation

To collect the news articles, we obtained access to the general news feed from Lusa. This feed provided a diverse range of articles covering various topics and events. In this first iteration, linguists analyzed a set of 670 news articles to determine if they were predominantly narrative. This process involved identifying specific features with the aim of selecting articles that prominently featured narratives.

Firstly, we imposed a length restriction on the selected articles to optimize the annotation process and reduce the workload. By limiting the length of the articles, we aimed to ensure a manageable annotation task while still capturing meaningful narrative content. For this reason, we decided that the total word count of each news should be between 50 and 200.

Secondly, based on the analysis of the 670 news, we also specified that news in English were to be discarded, as well as those that were very similar to each other. Subsequently, after this initial filtering, we conducted a search for the most frequently occurring words and expressions in the selected articles. This list included the following 37 Portuguese words: [*arma (gun), autoridades (authorities), caso (case), concelho (county), crime (crime), crimes (crimes), deteve (detained), detido (detainee), distrito (district), estado (state), gnr (gnr), governo (government), homem (man), judiciária (judiciary), lisboa (lisbon), morte (death), mortos (deaths), mulher (woman), ocorreu (happened), país (country), pessoas (people), pj (pj), polícia (police), polícia judiciária (judiciary police), presidente (president), segurança (security), suspeito (suspect), tribunal (court), vítima (victim)*]. These are words highly associated with crime-related news, which tend to be more narrative, as they recount stories involving related events such as robberies or illegal activities interrupted by authorities. From the initial set, these news articles proved to be the most suitable for creating a narrative news corpus. In addition to providing basic information about key news elements (Who?, What?, When?, Where?, How?, and Why?), these selected news articles included the description of events related to the main occurrence, enriching the narrative by detailing events that happened before or after.

The article selection process took place in January 2021, with Lusa using their internal system to search for the selected keywords. The linguistics team validated the search, and 357 news articles were selected to constitute the dataset. The list of articles was provided in a semi-structured format, which was then parsed to generate a single JSON file containing all the articles, structured as shown in Listing 1. The combination of the filtering criteria and length restriction resulted in a curated collection of news articles that specifically emphasize narratives.

3.2. Dataset Annotation

A subset of the Text2Story Lusa dataset, which we named Text2Story Lusa Annotated Corpus, was then used to develop and test an annotation scheme. This subset comprises 117 news articles that were annotated with semantic information. The annotation scheme was developed by the Text2Story linguistics team that combined and harmonized for the first time four different parts of the Standard Language Resource Management-Semantic Annotation Framework (Silvano et al., 2021b; Leal et al., 2022; Silvano et al., 2023): Part 1 - Time and events (ISO-24617-1, 2012), Part 4 - Semantic roles (ISO-24617-4, 2014), Part 7 - Space (ISO-24617-7, 2020), and Part 9 - Referential annotation framework (ISO-24617-9, 2019).

This multilayer semantic annotation scheme comprises entity structures (events, times, participants, measures, and spatial relations) and link structures (e.g., temporal, aspectual, subordination, objectal, spatial, and semantic role links). For each entity and link structure, several attributes were selected to characterize the structure. For instance, the entity structure for participants includes the following attributes (and values): lexical head (a noun or a pronoun); domain (a set of individuals, a single individual or a mass quantity); type (e.g., a person, an organization, a facility, an object); involvement (i.e., 0, 1, more than 1, all, undefined quantification).

```
{
  "articles": [
    {
      "id": "102",
      "location": "Lisboa",
      "publication_time": "2020-12-12",
      "headline": "Homem socorrido no Rio Tejo está livre de perigo",
      "content": "Um homem de 68 anos caiu hoje no Rio Tejo, junto ao Cais das Colunas, [...] onde ainda se encontra internado, mas livre de perigo."
    }
  ]
}
```

Listing 1: Text2Story Lusa news article example.

The primary purpose of the annotation scheme was to be able to represent the main morphosyntactic and semantic features of the relevant narrative components, that is the events (What, How, Why), the participants (Who), the time (When), the space (Where), and also the different relationships between them. The projection of the several entity and link structures, with several attributes that codify their grammatical characteristics, permits to accomplish that main objective. This first subset of data served as a case study for the linguists' team to design the annotation scheme incrementally. Therefore, it reflects an ongoing process of building a far-reaching narrative annotation framework. The linguistics team carried out the annotation manually and hand-validated at each step. After designing the scheme, we created the annotation guidelines.

Afterwards, group hands-on sessions were organised to discuss and resolve annotation challenges. Whenever required, adaptations to the scheme and the guidelines were introduced. The next phase was the individual annotation followed by meetings to determine the degree of agreement among team members. The necessary adjustments to the framework and guidelines were made after each iteration. Throughout this process, the team consolidated and validated the annotation framework.

The manual annotation was done using BRAT (brat rapid annotation tool) (Stenetorp et al., 2012), which facilitated updating the annotation scheme during its construction. Each news article has an associated annotated file in the standard BRAT format. Listing 2 illustrates this format, wherein each annotated entity is assigned an associated ID and attributes, and relations also have IDs and entities as arguments. For instance, [*a PSP* (the PSP)] is a participant of ID *T53* with the following attributes and values: lexical head: Noun (ID *A286*); domain: Individual (ID *A287*); type: Org (ID *A288*); involvement: 1 (ID *A289*). Additionally, [*a PSP*] is linked to the event [*Segundo* (According to)] (ID *T64*) by the semantic role of *Agent* (ID *R86*).

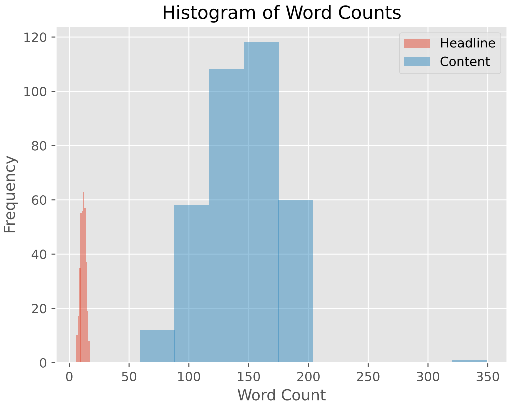

Figure 1: Word count histogram for the 'headline' and the 'content' fields.

3.3. Dataset Characterization

The Text2Story Lusa dataset (Nunes et al., 2023) is composed of 357 news articles, containing a total of 55,414 words (11,152 unique). The dataset's structure is shown in Listing 1. Each article is composed of the following fields: 'id' (number), 'location' (text), 'publication_time' (YYYY-MM-DD format), 'headline' (text), and 'content' (text). This section broadly characterizes the textual fields and reports on the results. In Figure 1, the length distribution of each field is depicted using a histogram. As expected, headlines exhibit much shorter lengths on average (11.52) compared to the body (143.70).

In Figure 2, the 20 most common words in both the headline and the content are listed. For tokenization and punctuation removal of the texts, the NLTK Python library was used (NLTK Team, 2023). The most common words in the title are related to the participants or the actors involved in the news, either identifying them (e.g., *homem* (man),

```
T64     Event 930 937 Segundo
A91     Class T64 Reporting
T53     Participant 938 943 a PSP
A286    Lexical_Head T53 Noun
A287    Individuation_Domain T53 Individual
A288    Participant_Type_Domain T53 Org
A289    Involvement T53 1
T54     Participant 945 953 o detido
A290    Lexical_Head T54 Noun
A291    Individuation_Domain T54 Individual
A292    Participant_Type_Domain T54 Per
A293    Involvement T54 1
T21     Time 956 960 hoje
A145    Time_Type T21 Date
A146    TemporalFunction T21 Publication_Time
T22     Event 961 996 presente às Autoridades Judiciárias
A147    Class T22 Occurrence
A148    Event_Type T22 Transition
A149    Pos T22 Verb
A150    Tense T22 Present
A154    Polarity T22 Pos
R86     SRLINK_agent Arg1:T53 Arg2:T64
R67     TLINK_after Arg1:T22 Arg2:T64
R78     SRLINK_patient Arg1:T22 Arg2:T54
R19     TLINK_isIncluded Arg1:T22 Arg2:T21
```

Listing 2: An annotation excerpt, in the BRAT standoff format, from the Text2Story Lusa Dataset. The complete annotated sentence is “Segundo a PSP, o detido é hoje presente às Autoridades Judiciárias.” (Translated: “According to the PSP, the detainee is present today at the Judicial Authorities”).

PJ, GNR (Police), *mulher* (woman)) or providing information about them (e.g., age). Some of the main words represent the events reported in the news, namely *suspeito* (suspect), *mortos* (dead), and *deteve* (arrested). Additionally, alongside words identifying participants, the most frequent words in the content portion include more events, as expected, some of which are related to the reporting layer (e.g., *disse* – ’said’). Interestingly, the most recurring word in this section of the news is *hoje* (today), which corresponds to the temporal location of the event, a key element in news reporting.

To analyze the distribution of lexical categories in the dataset, we used the spaCy Python library Part-of-Speech (POS) tagging tool (Explosion AI, 2023). Figure 3 presents the most frequent POS types. These results align with the ones for the most common words: participants are primarily represented by nouns, while events are represented by verbs. Therefore, both in the title and the content, nouns, both common and proper, are in higher percentages. Adpositions, which include prepositions introducing, for example, temporal and spatial phrases or participants, also occur in great numbers. Since news report on situations, describing features of participants or places, as well, verbs are expected to surface frequently.

As previously described, on top of the Text2Story Lusa dataset, we added a multilayer annotation to a part of the documents which resulted in the Text2Story Lusa Annotated Corpus dataset (Silvano et al., 2023). This collection comprises a subset of 117 news articles fully annotated following the multilayer scheme that was described in the previous section. The frequencies of the elements described in the scheme are portrayed in Table 1. First, these numbers show the scale of the annotation effort, with over 51 thousand annotated elements and attributes – encompassing 7,516 entity annotations, 12,009 link annotations, and 31,129 defined attributes. Additionally, they show that the semantic role links are the most significant in number, as they indicate the involvement of participants in events. Objectal links also demonstrate a high number, as they establish relationships between entities referred to in the texts. Temporal links and qualitative spatial links also exhibit considerable numbers, as they provide temporal and spatial information. These four types of links are recurrent due to the nature of the news, as they typically provide information on who did what, where, and when.

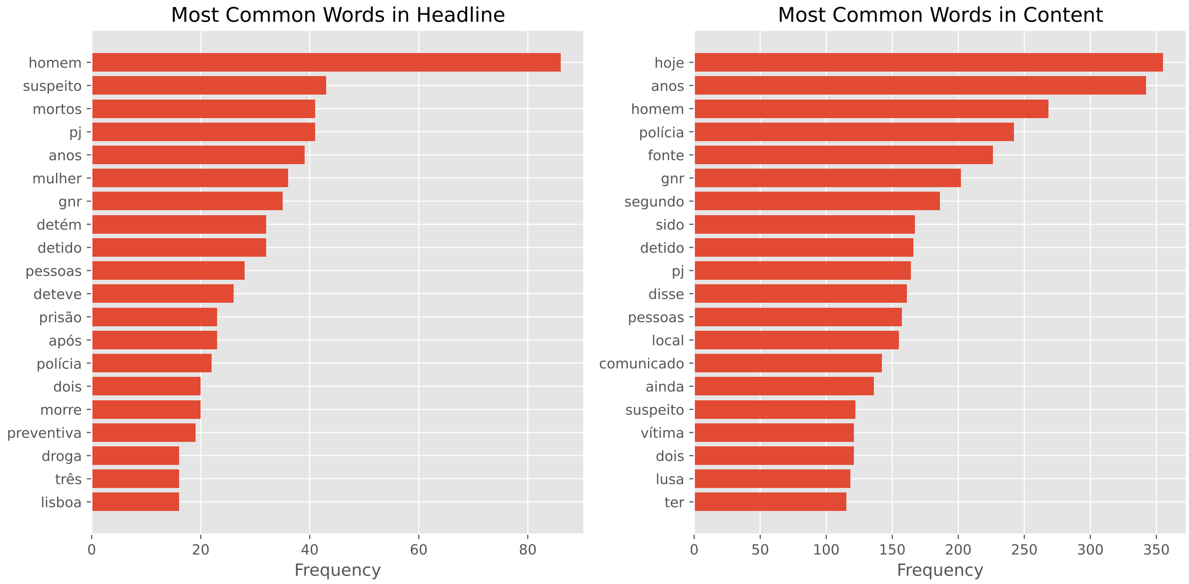

Figure 2: Most common words in the 'headline' and the 'content' fields.

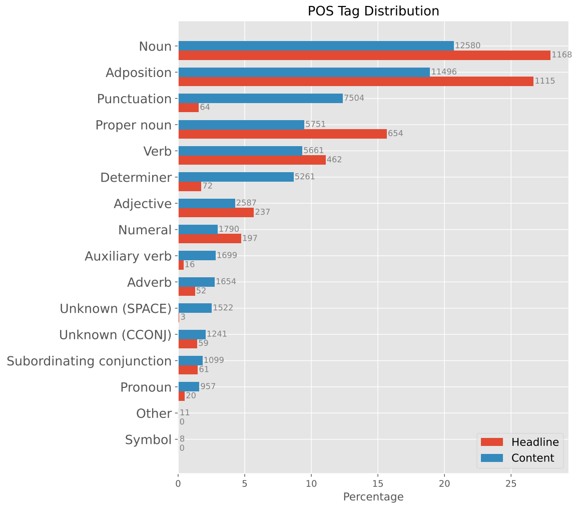

Figure 3: Distribution of POS types in the 'headline' and the 'content' fields.

4. Usage

As stated, the main purpose of the Text2Story Lusa datasets is to help the development of methods for narrative extraction and representation. The datasets have already been used and are actively exploited for building and testing narrative extraction tools. In one of the projects, Brat2Viz (Amorim et al., 2021), the authors implemented a tool that transforms texts annotated with our framework into a formal notation (Discourse Representation Structure) and subsequently into visual representations (knowledge graphs and message sequence charts). Part of the Text2Story Lusa Annotated Corpus dataset is used to showcase this tool[6]. This dataset has also been used for the development of a specific form of visualization within the scope of the Brat2Viz tool. In this case, the annotated narratives were automatically represented using icons (Valente et al., 2022).

Narrative Component	Frequency
Participants	3,530
Events	3,027
Spacial Relations	512
Times	438
Measure	9
Semantic Role Links	4,764
Temporal Links	3,522
Objectal Links	2,224
Qualitative Spatial Link	927
Subordination Link	429
Movement Link	125
Aspectual Link	18

Table 1: Frequency of each narrative component in the Text2Story annotated corpus.

The Text2Story Lusa Annotated Corpus dataset was critical to studying the role of reporting events in this type of narrative, some of their lexicon-grammatical features, and how they related to the events that compose the story (Silvano et al., 2023).

This dataset is also being used for routinely testing a narrative extraction pipeline[7] on different tasks, primarily focused on the identification of par-

[6] https://nabu.dcc.fc.up.pt/brat2viz

[7] https://pypi.org/project/text2story/

ticipants, events, temporal expressions, and the relations between these elements. The performance of the most popular Large Language Models has been assessed on these challenging tasks using a prompting approach. While LLMs have shown good results, they are still far from perfect (Sousa et al., 2023). The authors have systematically explored various prompting configurations to enhance performance.

5. Discussion

This section addresses ethical considerations and discusses the principles of making this resource available under FAIR (Findable, Accessible, Interoperable, and Reusable) principles.

Overall, the Text2Story Lusa datasets include news articles that are primarily focused on crime events, as indicated by the selected keywords. This enhances the narrative structure of the collection, as this kind of text embodies a wealth of narrative elements. However, the specialized nature of the dataset, with its focus on crime events, may limit its applicability in broader tasks or domains. Relying predominantly on crime narratives might introduce biases when training machine learning models, such as language models or classification algorithms. For instance, using this dataset for general language modeling could result in models that disproportionately represent crime-related terms and concepts, potentially skewing outcomes and perceptions when applied to broader tasks. It's essential for researchers to be aware of this specificity and be cautious when considering the dataset for applications outside its primary scope.

In line with the FAIR principles (Wilkinson et al., 2016), we have undertaken several steps to ensure the dataset's accessibility and reusability. First, the dataset is published under a Creative Commons Attribution-NonCommercial license, allowing researchers and practitioners to access and utilize it for non-commercial purposes. We have also adhered to a precise protocol established with the intellectual property holders, Lusa, to ensure legal and ethical considerations.

We made efforts to enhance the findability and accessibility of the dataset. The Text2Story Lusa datasets are shared through a dataset-sharing service that indexes both the dataset and metadata, providing a DOI for easy citation and discovery. Furthermore, each dataset is accompanied by clear documentation and a description of its contents, facilitating its use and interpretation. Finally, we also address interoperability issues. At the representation level by providing the dataset through standard structured formats, specifically JSON and BRAT. Moreover, the adoption of ISO standards for annotation promotes interoperability not just at the representation level but also at the annotation level, thus enabling adopting by other researchers.

6. Conclusions and Future Work

In this paper, we presented Text2Story Lusa, a linguistic resource comprised of two datasets developed to support research in various fields, such as linguistics and computer science. Text2Story Lusa includes a manually curated collection of news articles written in European Portuguese, obtained from the Lusa news agency. Through a targeted dataset creation process, we have focused on selecting articles that prominently feature narratives. Text2Story Lusa Annotated Corpus includes manual annotations for a subset of the original articles. The annotation adopts a multilayer semantic scheme comprising entity structures (events, times, participants, measures, and spatial relations) and link structures (e.g., temporal, aspectual, subordination, objectal, spatial, and semantic role links).

These resources fill an important gap in language resources for European Portuguese, particularly in the domain of narrative extraction and semantic analysis. Researchers in the fields of Natural Language Processing, Linguistics and related disciplines can leverage them for various tasks and challenges beyond narrative extraction.

In future work, we plan to expand these resources by including more extensive news, covering a wider range of topics and genres, and providing additional annotations. Furthermore, we will explore avenues for integrating the Text2Story Lusa dataset with other existing resources to enhance cross-domain research possibilities (e.g., entity linking with existing open knowledge bases).

In conclusion, the Text2Story Lusa datasets open up new opportunities for research and development, addressing the growing need for high-quality language resources in European Portuguese.

7. Acknowledgements

We acknowledge and thank the collaboration of Lusa - Agência de Notícias de Portugal, S.A., for providing access to this collection, greatly contributing to the development of language resources for European Portuguese. In particular, we are deeply grateful for the central contribution of the late Pedro Camacho, whose pivotal role in establishing the collaboration with Lusa was fundamental to the success of this project.

This work was financed by national funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P. (Portuguese Foundation for Science and Technology) within the project StorySense, with reference 2022.09312.PTDC.

8. Bibliographical References

Evelin Amorim, Alexandre Ribeiro, Brenda Salenave Santana, Inês Cantante, Alípio Jorge, Sérgio Nunes, Purificação Silvano, Antonio Leal, and Ricardo Campos. 2021. Brat2viz: a tool and pipeline for visualizing narratives from annotated texts. In *Proceedings of Text2Story - Fourth Workshop on Narrative Extraction From Texts held in conjunction with the 43rd European Conference on Information Retrieval (ECIR 2021), Lucca, Italy, April 1, 2021 (online event due to Covid-19 outbreak)*, volume 2860 of *CEUR Workshop Proceedings*, pages 49–56. CEUR-WS.org.

Florbela Barreto, António Branco, Eduardo Ferreira, Amália Mendes, Maria Fernanda Bacelar do Nascimento, Filipe Nunes, and João Ricardo Silva. 2006. Open resources and tools for the shallow processing of Portuguese: The TagShare project. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Valerio Basile, Johan Bos, Kilian Evang, and Noortje Venhuizen. 2012. Developing a large semantically annotated corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3196–3200, Istanbul, Turkey. European Language Resources Association (ELRA).

Johan Bos, Valerio Basile, Kilian Evang, Noortje Venhuizen, and Johannes Bjerva. 2017. *The Groningen Meaning Bank*, pages 463–496.

Paula Cardoso, Erick Maziero, Mara Luca Castro Jorge, Eloize Seno, Ariani Di Felippo, Lucia Helena Machado Rino, Maria das Gracas Volpe Nunes, and Thiago Pardo. 2011. CSTnews-a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In *Summarization of News Texts in Brazilian Portuguese. In the Proceedings of the 3rd RST Brazilian Meeting*.

Sandra Collovini, Thiago Carbonel, Juliana Fuchs, Jorge Coelho, Lúcia Rino, and Renata Vieira. 2007. Summ-it: Um corpus anotado com informações discursivas visando à sumarização automática. In *V Workshop em Tecnologia da Informação e da Linguagem Humana – TIL. Proceedings of XXVII Congresso da SBC*, Rio de Janeiro.

Explosion AI. 2023. spaCy - Industrial-strength Natural Language Processing in Python. Version 3.5.2.

Michel Généreux, Iris Hendrickx, and Amália Mendes. 2012. Introducing the reference corpus of contemporary portuguese online. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.

Nancy Ide. 2017. *Case Study: The Manually Annotated Sub-Corpus*, pages 497–519. Springer Netherlands, Dordrecht.

Nancy Ide, Collin Baker, Christiane Fellbaum, Charles Fillmore, and Rebecca Passonneau. 2008. MASC: the manually annotated subcorpus of American English. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Nancy Ide, Laurent Romary, and Eric de la Clergerie. 2003. International standard for a linguistic annotation framework. In *Proceedings of the HLT-NAACL 2003 Workshop on Software Engineering and Architecture of Language Technology Systems (SEALTS)*, pages 25–30.

ISO-24617-1. 2012. Language resource management- Semantic annotation framework (SemAF) - Part 1: Time and events (SemAF-Time, ISO-TimeML). Standard, International Organization for Standardization, Geneva, CH.

ISO-24617-4. 2014. Language resource management- Semantic annotation framework (SemAF) - Part 4: Semantic roles (SemAF-SR). Standard, International Organization for Standardization, Geneva, CH.

ISO-24617-6. 2016. Language resource management- Semantic annotation framework (SemAF) - Part 6:Principles of semantic annotation (SemAF Principles). Standard, International Organization for Standardization, Geneva, CH.

ISO-24617-7. 2020. Language resource management-Semantic annotation framework (SemAF) - Part 7: Spatial information. Standard, International Organization for Standardization, Geneva, CH.

ISO-24617-9. 2019. Language resource management- Semantic annotation framework (SemAF) - - Part 9: Reference annotation framework (RAF). Standard, International Organization for Standardization, Geneva, CH.

Alípio M. Jorge, Ricardo Campos, Adam Jatowt, and Sérgio Nunes. 2019. Information Processing & Management Journal Special Issue on Narrative Extraction from Texts (Text2Story): Preface. *Information Processing & Management*, 56(5):1771–1774.

António Leal, Purificação Silvano, Evelin Amorim, Inês Cantante, Fátima Silva, Alípio Mario Jorge, and Ricardo Campos. 2022. The place of ISO-Space in Text2Story multilayer annotation scheme. In *Proceedings of the 18th Joint ACL - ISO Workshop on Interoperable Semantic Annotation within LREC2022*, pages 61–70, Marseille, France. European Language Resources Association.

NLTK Team. 2023. NLTK :: Natural Language Toolkit. Version 3.8.1.

Paulo Alexandre Rocha and Diana Santos. 2000. CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In *Actas do V Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada, PROPOR'2000. 19-22 de Novembro de 2000.*, Atibaia, São Paulo. ICMC/USP.

Brenda Santana, Ricardo Campos, Evelin Amorim, Alípio Jorge, Purificação Silvano, and Sérgio Nunes. 2023. A survey on narrative extraction from textual data. *Artif. Intell. Rev.*, 56(8):8393–8435.

Purificação Silvano, Evelin Amorim, Antonio Leal, Inês Cantante, Fátima Silva, Alípio Jorge, Ricardo Campos, and Sérgio Nunes. 2023. Annotation and visualisation of reporting events in textual narratives. In *Proceedings of Text2Story - Sixth Workshop on Narrative Extraction From Texts held in conjunction with the 45th European Conference on Information Retrieval (ECIR 2023), Dublin, Ireland, April 2, 2023*, volume 3370 of *CEUR Workshop Proceedings*, pages 47–62. CEUR-WS.org.

Purificação Silvano, António Leal, Fátima Silva, Inês Cantante, Fátima Oliveira, and Alípio Mário Jorge. 2021a. Developing a multilayer semantic annotation scheme based on ISO standards for the visualization of a newswire corpus. In *Proceedings of the 17th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 1–13.

Purificação Silvano, António Leal, Fátima Silva, Inês Cantante, Fatima Oliveira, and Alípio Mario Jorge. 2021b. Developing a multilayer semantic annotation scheme based on ISO standards for the visualization of a newswire corpus. In *Proceedings of the 17th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 1–13, Groningen, The Netherlands (online). Association for Computational Linguistics.

Hugo Sousa, Nuno Guimarães, Alípio Jorge, and Ricardo Campos. 2023. GPT struct me: Probing GPT models on narrative entity extraction. In *IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2023, Venice, Italy, October 26-29, 2023*, pages 1–8. IEEE.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Fiona Anting Tan, Ali Hürriyetoğlu, Tommaso Caselli, Nelleke Oostdijk, Tadashi Nomoto, Hansi Hettiarachchi, Iqra Ameer, Onur Uca, Farhana Ferdousi Liza, and Tiancheng Hu. 2022. The causal news corpus: Annotating causal relations in event sentences from news. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2298–2310, Marseille, France. European Language Resources Association.

Joana Valente, Alípio Jorge, and Sérgio Nunes. 2022. Text2icons: linking icons to narrative participants (position paper). In *Proceedings of Text2Story - Fifth Workshop on Narrative Extraction From Texts held in conjunction with the 44th European Conference on Information Retrieval (ECIR 2022), Stavanger, Norway, April 10, 2022*, volume 3117 of *CEUR Workshop Proceedings*, pages 111–116. CEUR-WS.org.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Amir Zeldes and Dan Simonson. 2016. Different flavors of GUM: Evaluating genre and sentence type effects on multilayer corpus annotation quality. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 68–78, Berlin, Germany. Association for Computational Linguistics.

9. Language Resource References

Amália Mendes and Iris Hendrickx. 2021. *News corpus categorised [Dataset]*. CLUL. Portulan Clarin.

n.a. 2016. *FEUP news corpus [Dataset]*. LIACC. Portulan Clarin.

Sérgio Nunes and Alípio Jorge and António Leal and Evelin Amorim and Hugo Sousa and Inês Cantante and Purificação Silvano and Ricardo Campos. 2023. *Text2Story Lusa [Dataset]*. INESC TEC and CLUP. INESC TEC.

Purificação Silvano and Alípio Jorge and António Leal and Evelin Amorim and Hugo Sousa and Inês Cantante and Ricardo Campos and Sérgio Nunes. 2023. *Text2Story Lusa Annotated Corpus [Dataset]*. INESC TEC and CLUP. INESC TEC.